基于混合云架构的深度语义密文检索

李　剑，矫　健

(北京邮电大学 人工智能学院，北京 100876)

摘　要：针对传统的云环境下密文检索方案基于统计学模型来生成文件向量和检索向量，并没有考虑文件和请求的深层次语义信息，提出一种基于混合云架构的深层次语义密文检索模型。通过私有云联邦学习神经网络模型构建向量生成模型，通过公有云存储密文数据。另外，提出密倒排索引表来存放文件向量，在公有云的检索过程中，保证检索信息不被泄露的情况下提高检索的效率。对真实数据集的分析和实验表明，提出的方案在安全性和搜索效率方面都优于目前同类型的密文检索方案。

关键词：密文检索；混合云；联邦学习；加密倒排索引表

中图分类号：TP309　　doi: 10.19734/j.issn.1001-3695.2022.02.0101

Deep semantic ciphertext retrieval based on hybrid cloud architecture

Li Jian, Jiao Jian

(*School of Artificial Intelligence*, *Beijing University of Posts & Telecommunications*, *Beijing* 100876, *China*)

Abstract: Aiming at the traditional ciphertext retrieval scheme in cloud environment, which generates file vectors and retrieval vectors based on statistical model, and does not consider the deep-seated semantic information of files and requests, this paper proposes a deep-seated semantic ciphertext retrieval model based on hybrid Cloud Architecture. The vector generation model is constructed through the private cloud federated learning neural network model, and the ciphertext data is stored through the public cloud. In addition, this paper proposes a secret inverted index table to store file vectors, so as to improve the efficiency of retrieval without ensuring that the retrieval information is not leaked in the retrieval process of public cloud. The analysis and experiments on real data sets show that this scheme is better than the current ciphertext retrieval schemes of the same type in terms of security and search efficiency.

Key words: ciphertext retrieval; hybrid cloud; federal learning; encrypted inverted index table

0　引言

随着网络技术的发展和网络业务需求的扩大，以及云计算和大数据的发展。为了提高效率，越来越多的机构和公司将数据上传至公有云服务器。然而数据隐私一直是云计算应用发展的一个重大阻碍。虽然云服务提供商声称防火墙等机制可以增强用户数据的安全性，但公有云服务器完全控制外包数据，“诚实但好奇的”的云服务器可能会泄露数据所有者不愿透露的敏感数据。因此，数据所有者在将文档上传到半诚实的云之前对其进行加密，并将数据存储为密文以确保文档的安全性。另一方面，企业和政府为保证数据安全性构建私有云服务器，往往存在内部的大量有价值资源不可被使用，出现数据孤岛问题。因此提出一种在云环境下通过数据共享提升性能的高效密文检索方案具有重要意义。

数据加密方案对数据检索提出了巨大的挑战。近年来，研究学者提出了许多文本检索策略。在文献[1]中，Cao 等人首先采用向量空间模型计算文档向量和查询向量的内积，并使用安全 kNN 算法(Sec-kNN)对其进行加密。基于内积运算的结果，提出了一种多关键字密文检索结果排序(MRSE)方案。后来，研究者提出了许多拓展方法[2~5]。这些方法都具有可证明的安全性，但在信息检索领域都采用了传统的 TF-IDF 加权统计计算规则。使用关键词的规则无法有效捕获单词的上下文。此外，这些方案具有高向量维数、高存储要求和高时间复杂度。在信息检索中，当单词匹配失败时，潜在语义模型将查询映射到相关文档。该模型解决了文档和查询之间的语言差异。具体来说，即使查询关键词没有在文档中直接出现，它们也可能被构造为两个具有高相似度的低维语义向量。

语义搜索是明文数据和加密数据信息检索的一个重要研究方向[6~11]。语义分析消除了查询和文档之间的语言差异。Fu 等人[12]在语义本体 的支持下建立了用户兴趣模型，实现了个性化的关键词精确搜索。在 文献[13,14]中，使用互信息模型构建语义扩展方案。例如，Jadhav 利用互信息模型扩展查询关键字，然后计算文档的相关性得分。文献[15]中提出了模糊关键词搜索技术，用于扩展关键词集合。Fu 等人[15]开发了基于单纯形的分级多关键字模糊搜索方案，没有预定义的模糊集。Yang 等人[16]提出在一个可验证的语义方案中利用 EMD 距离，该语义方案描述了查询和文档之间的单词传输问题，单词传输的最小成本称为查询和每个文档之间的相似性分数。

本文使用混合云架构，解决数据孤岛问题，利用联邦学习构建深度神经网络模型提取数据深层语义，并提出加密倒排索引表结构来缩短检索时间，该方案保证了数据的安全性，并有较高的检索准确性和效率。

1　问题描述

1.1　系统模型

如图 1 所示，在本方案中，一共有五个实体，分别为数据拥有者、数据使用者、私有云服务器、公有云服务器、参数服务器。

1)数据拥有者

数据拥有者拥有有价值的数据。对数据进行加密处理后

收稿日期：2022-02-09；**修回日期：**2022-04-24

作者简介：李剑(1976-)，男，陕西西安人，教授，博导，博士，主要研究方向为信息安全、量子密码学；矫健(1996-)，男，辽宁大连人，硕士研究生，主要研究方向为云计算安全(joe@bupt.edu.cn).

上传至公有云，同时将明文数据上传至私有云后进行模型训练。根据训练完成的神经网络模型，生成文件向量后，构建加密倒排索引表上传至公有云。

2)数据使用者

数据使用者将检索关键词、个人信息发送至所有数据拥有者进行授权认证、检索关键词映射字及密钥。将检索关键词发送至参数服务器接收到检索向量后，根据密钥生成加密陷门发送至公有云，并收到公有云返回的 TOP-K 相关文件结果。

3)私有云服务器

每个数据拥有者有一个私有云服务器。“诚实且可信”的私有云服务器收到明文数据、网络模型后单独训练，每轮训练结束后将参数结果与参数服务器进行联邦训练。训练好的神经网络模型回传至数据拥有者。

4)公有云服务器

公有云服务器为“诚实且不可信”的实体。存储数据拥有者的加密数据与加密倒排索引表。收到参数服务器发送的请求陷门后，通过计算找出相关加密文件发送至数据使用者。

5)参数服务器

参数服务器同样也是私有云服务器。是“诚实且可信”的，由所有数据拥有者共同维护。它是联邦学习模型训练的中央服务器。在收到数据使用者的检索关键字后生成检索向量发送至数据使用者。

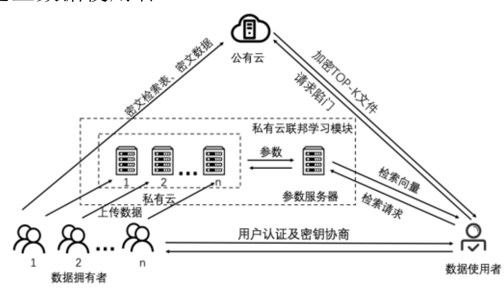


图 1　系统模型

Fig. 1　System model

1.2　威胁模型

在本文的方案中，本文假设参数服务器是可信的，公共云是一个“诚实但好奇”的服务器[2]。基于半诚实公共云服务器已知的信息，本文研究了两种威胁模型。

已知密文威胁模型。公共云只知道加密文档、加密数据索引和安全查询陷门。在这种情况下，公共云服务器仅使用仅密文攻击模式进行攻击。

已知背景威胁模型。公共云服务器应该知道比已知密文模型中更多的信息。这些知识包括陷门的相关关系，以及与数据集相关的统计信息。公共服务器使用已知的陷门信息来分析查询关键字或陷门与文档的关系。

1.3　符号

Q:检索关键词集合， $Q=\{q_1,q_2,\ldots,q_k\}$ 。

D:明文数据集合 $D=\{D^1,D^2,\ldots,D^n\}$ ，其中 D^i 代表第 i 个数据拥有者的明文数据集合。

E:密文数据集合 $E=\{E^1,E^2,\ldots,E^n\}$ ，其中 E^i 代表第 i 个数据拥有者的密文数据集合 i 。

P:文件向量集合， $P=\{P_1^1,P_2^1,\ldots,P_j^n\}$ ，其中 P_t^i 代表第 i 个数据拥有者的第 t 个明文数据对应的文件向量。

I:加密文件向量集合， $I=\{I_1^1,I_2^1,\ldots,I_j^n\}$ ，其中 I_t^i 代表第 i 个数据拥有者的第 t 个明文数据对应的加密文件向量。

N : 检 索 关 键 词 映 射 数 字 集 合 ， $N=\{N_1^1,N_2^1,\ldots,N_k^n\}$ ，其中 N_t^i 代表第 t 个检索关键词在第 i 个数据拥有者的映射数字表达。

V:检索向量。

T:加密检索向量。

1.4　设计目标

为了在上述模型下有效地利用外包的云数据实现排序搜索，本文方案需要确保检索准确性和检索效率。

检索准确率：使用统计特征的经典加密检索模型无法捕获单词的上下文信息和文档的深层语义。本文的方案旨在研究文件深层次的语义，而不是基于统计学特征作为文件检索结果的依据。，本文的模型不是使用统计特征的方法，而是通过神经网络进行优化，检索准确率高于以前的方法。

效率：效率包括两个方面，搜索和存储。减小生成向量的维数可天然减小存储与计算资源消耗，同时设计合适的文档索引向量管理结构也可提高检索的效率。

2　模型

2.1　私有云联邦学习模型

联邦学习是分布式机器学习架构的一种表现形式，联邦学习可分为训练服务器和中心参数服务器。所有服务器共享需要训练的机器学习模型，且共享各服务器每轮训练的参数，所有训练服务器将参数以密文形式传递给中心服务器后，中心服务器进行参数统一。联邦学习架构，在保证数据不被共享的前提下，对所有数据进行集中训练。解决了各个数据拥有者的敏感数据孤岛问题。

联邦学习由参数服务器与数据训练方两部分组成，所有训练者共享训练模型[17]。各数据训练者单独训练自己的数据，共享训练参数。传统密文检索方案基于统计学模型，根据文件中关键词的词频与逆向文件频率来生成文件向量和检索请求向量。在此基础上提高检索准确性的方案为检索关键词的拓展，如用户兴趣模型、根据深度学习得出的相似关键词的拓展，或根据关键词的位置进行权重更新。由于数据安全性的问题，并没有通过深度神经网络训练明文数据，挖掘文章深层次语义。本文提出基于私有云架构下的联邦学习模型，可将文章向量与检索向量的生成方式由传动的统计学模型更新为深度学习模型。并且考虑所有数据训练者的计算性能有所不同，设计时间窗口管理模式，提高了通信与训练效率。如图 2 所示，为私有云联邦学习模型架构。

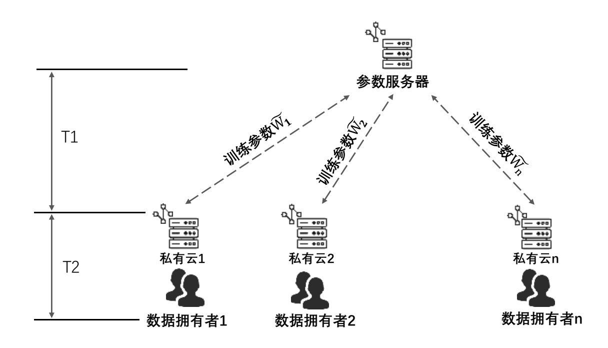


图 2　私有云联邦学习模型

Fig. 2　Private cloud federation learning model

一轮联邦学习网络模型更新的时间由数据拥有者的训练时间 T1 和参数传递及参数服务器更新时间 T2 组成。考虑到每个数据拥有者的数据量及计算能力不同，本文设计了时间窗口管理模型。联邦学习开始之间，整个系统测试通信时间，设定通信窗口 T1 及总体窗口时间 T，数据拥有者训练时间 T2=T-T1。对数据拥有者 1 来说，第一次训练参数结果为 W_1 ，后在 T2 时间内继续本地训练，本地训练的轮次根据数据拥有者的数据量大小、私有云的计算能力不同也会有所不同，在结束时的训练参数结果计为 $\widetilde{W_1}$ ，将结果发送至参数服务器。该方案既保证了每轮模型统一时，所有用户数据拥有者都参与训练又考虑了计算能力强的数据拥有者可多次本地训练提升模型训练效果。

2.2　神经网络向量生成模型

与传统的统计学生成的向量生成模型不同，本文利用私有云联邦学习模型训练神经网络向量生成模型。模型选择为 DSSM[18]。文件向量、检索向量通过该模型映射到同维度的深层次语义空间。

DSSM 模型的输入为 N 个文件与 1 个查询请求。神经网络架构为五层。第一层为维度为 500k 的输入。经过 Word-n-gram 层将维度减小到 30k。后经过两层全连接深度神经网络层，维度为 300、300，后输入为 128 维的向量。该模型的各层激活函数为 tanh。

在训练过程中，通过该模型生成的 N+1 个 128 位的向量对应了文件与查询请求。为简化密文检索的复杂度，本文在生成 128 维向量后，需对向量进行归一化处理：

$$y=\frac{Y}{\|Y\|} \tag{1}$$

相关性分数为查询向量与文件向量的点乘结果。模型目标是优化点击文档的可能性，损失函数如式(2)所示。

$$L(\Lambda)=-\log\prod_{(Q,\bar{D}^{+})}P(\bar{D}^{+}\mid Q) \tag{2}$$

其中，通过 softmax 函数计算的正向文档与查询请求的相关性得分的后验概率，Υ 为以真实数据为背景得出的平滑系数，$\bar{D}$ 为所有文档，包括 4 个没有被点击的文档 $\bar{D}^{-}$ 和 1 个被点击的文档 $\bar{D}^{+}$。在本文的方案中，随机初始化参数，使用随机梯度算法使得每个私有云来分布式地训练。如式(3)所示。

$$P(\bar{D}^{+}\mid Q)=\frac{\exp(\Upsilon R(\bar{D}^{+}\mid Q))}{\sum_{D'\in D}\exp(\Upsilon R(\bar{D}'\mid Q))} \tag{3}$$

在本文中，每个数据拥有通过私有云服务器单独训练 DSSM 模型，数据为该拥有者自己的数据。通过参数服务器进行参数共享。

2.3　加密倒排索引表

倒排索引表在信息检索领域中应用广泛，根据 KEY-VALUE 格式建表，其中 KEY 为文件关键词集合。VALUE 值为文件的文件向量。在密文检索中，不可信的公有云存放检索表，避免公有云服务器对数据使用者的检索情况进行分析，所以不可将 KEY 值以明文关键字的形式直接存放于公有云服务器。本文利用离散对数难题将关键词映射为无规律的数字。并且在每轮检索过程中更新密文倒排索引表的 KEY 值。加密倒排索引表保证了数据和数据使用者的安全性，具体流程如图 3 所示。

数据使用者在检索的初始阶段需要向所有数据拥有者发送身份认证。认证通过后，数据使用者拥有 n 个密钥集合 $S=\{PD_1,q_1,a_1\},\{PD_2,q_2,a_2\},\ldots,\{PD_n,q_n,a_n\}$，其中 $\{PD_i,q_i,a_i\}$ 代表第 i 个数据拥有者与数据使用者共享的素数 q、整数 a，以及数据使用者 i 的文件密钥。

数据使用者通过私钥 X_{user} 生成公钥 Y_{user}，如式(4)所示。

$$Y_{user}=a^{(X_{user})}\bmod q \tag{4}$$

随后将 Y_{user} 和检索关键词集合 $Q=\{q_1,q_2,\ldots,q_k\}$ 一同发送给数据拥有者 U_i。数据使用者该轮请求有 k 个关键字。对于数据拥有者 U_i，根据 k 个关键字，生成 k 个私钥 $\{X_{pcs}^{1},\ldots,X_{pcs}^{k}\}$ 和 k 个公钥 $\{Y_{pcs}^{1},\ldots,Y_{pcs}^{k}\}$。并将公钥发送至数据使用者。

数据使用者通过式(5)计算出 k 个关键字映射数字：

$$N^{K}=(Y_{pcs}^{k})^{X_{user}}\bmod q \tag{5}$$

数据拥有者通过式(6)计算出 k 个关键字映射数字：

$$N^{K}=(Y_{user})^{X_{pcs}^{k}}\bmod q \tag{6}$$

在数据使用者与所有数据拥有者进行身份认证和关键字数字映射后，共有 $n\times k$ 个关键词映射数字生成，数据使用者将 $n\times k$ 个映射数字与检索关键词发送至参数服务器。每个数据使用者跟据自己拥有的 k 个映射数字更新加密倒排索引表对应 KEY 值，其余 KEY 值随机生成并发送至公有云。

3　本文方案

3.1　具体方案

$$key(1^{l(n)})\rightarrow\{SK,K,a,q\}$$

数据所有者通过随机密钥生成算法输出密钥以加密文档和索引，并输入一个安全参数 l。SK 是一个密钥集，包括一个 $(n+u+1)$ 位向量和两个 $(n+u+1)\times(n+u+1)$ 可逆矩阵，$SK=\{M_1,M_2,S\}$。此外，K 是对称密钥，a 是 q 的本源根。

$$\{a,q,X_{user},Y_{user}\}\rightarrow N$$

身份认证后，数据使用者发送检索关键字给所有数据拥有者，每个数据拥有者与该数据使用者根据 a，q 和各自的密钥 X_{user}，Y_{user} 来生成检索关键词的映射数字。具体流程如 2.2 所述。数据拥有者根据生成的关键字映射数字生成加密倒排索引表的 KEY 值。数据使用者拥有 $n\times k$ 个关键字映射数字。其中 k 为数据使用者的关键字个数，n 为数据拥有者的数量。

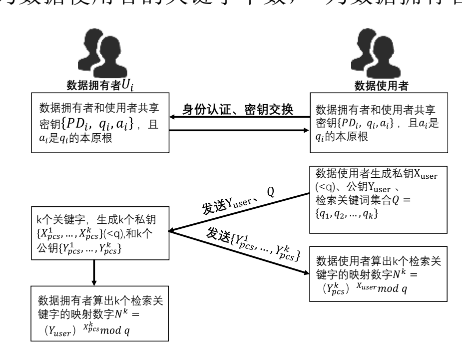


图 3　关键字映射数字过程

Fig. 3　Keyword mapping digital process

a)联邦学习深度神经网络模型

数据拥有者将数据信息传入私有云，共同训练预先规定好的神经网络模型。规定，在初始轮次神经网络的参数统一，在每轮训练结束、下一轮训练开始前，各私有云服务器从参数服务器下载最新参数，参数再次统一。在本文使用的神经网络模型为 DSSM[18]，该模型结构如 2.2 所述。联邦学习训练模型方式根据 2.1 所述，考虑所有用户参与每轮训练，也同时保证算力大的用户多次训练来提升联邦学习训练效率。

$$Index(SK,D)\rightarrow\{P,I\}$$

通过私有云训练好的向量模型，所有数据拥有者根据该模型明文数据 D 生成 128 维的文件向量 P，并将其扩展为 $(128+u+1)$ 维度向量 $\bar{P}$，其中扩展的最后一位设为 1，其他设为随机数字 $\varepsilon^{(j)}$。经过密钥 SK 加密文件向量 P 生成加密文件向量 I，具体为利用 S 将向量 $\bar{P}$ 分裂为 P' 和 P''，分裂规则如式(7)所示。数据使用者构建倒排索引表并将加密文件向量 I 放入表的 VALUE 位置。上传至公有云。

$$\begin{cases}\bar{P}_i'[t]=\bar{P}_i''[t]=\bar{P}[t] & (S[t]=0)\\ \bar{P}_i'[t]+\bar{P}_i''[t]=\bar{P}[t] & (S[t]=1)\end{cases} \tag{7}$$

$$Enc(D,K)\rightarrow C$$

数据使用者使用密钥 K 加密数据集 D。

$$Trapdoor(Q,SK)\rightarrow T$$

数据使用者将检索关键词发送至参数服务器，参数服务器根据训练好的模型生成检索向量发送至数据使用者，数据使用者根据密钥 SK 和接收到的 128 维检索向量 V，并扩展为 $(128+u+1)$ 维度向量 $\bar{V}$，扩展的最后一位为随机数字 t，其余补充位置由 0 或者 1 补充，并利用 S 将向量 $\bar{V}$ 分裂为 V' 和 V''，分裂规则如式(8)所示。再生成加密检索陷门 T 后，数据使用者将陷门 T 和关键词映射集合 N 发送至公有云。

$$\begin{cases}\bar{V}'[t]=\bar{V}''[t]=\bar{V}[t] & (S[t]=1)\\ \bar{V}'[t]+\bar{V}''[t]=\bar{V}[t] & (S[t]=0)\end{cases} \tag{8}$$

$$Search(T,I,N,k_{top})\rightarrow E_q$$

公有云收到数据使用者上传的加密检索陷门 T 后，根据关键词映射集合 N 与加密倒排索引表比对，找出与关键词映射集合 N 为 KEY 值的对应所有 VALUE 中的加密文件向量 I，通过式(9)计算出前 k_{top} 个文件 E_q，并发送至数据使用者。

$$I_i \cdot T = \{M_1^T \vec{P}_i, M_2^T \vec{P}_i'\}\cdot\{M_1^{-1}V', M_2^{-1}V''\} = \vec{P}_i \cdot V' + \vec{P}_i' \cdot V'' = \vec{P}\cdot\vec{Q} = r(P\cdot Q + \sum \varepsilon^{(j)}) + t \quad (9)$$

$$Dec(E_q, K) \rightarrow D_q$$

数据使用者根据收到结果文件集 E_q 后，根据密钥 K 解密后得到结果文件明文信息。

3.2 方案流程

本文提出的混合云架构的密文检索方案具体工作流程解耦为为两部分，非检索阶段和检索阶段，如图 4 所示。

非检索阶段包括构建向量生成模型，向量生成模型可生成检索向量和文件向量，通过章节 2.1 提出的混合云架构的联邦学习模型来训练 2.2 章节的神经网络模型，生成神经网络向量生成模型；每个数据拥有者拥有该模型输入明文数据得到文件向量，文件向量通过加密并利用 2.3 章节提出的加密倒排索引表来管理文件向量。

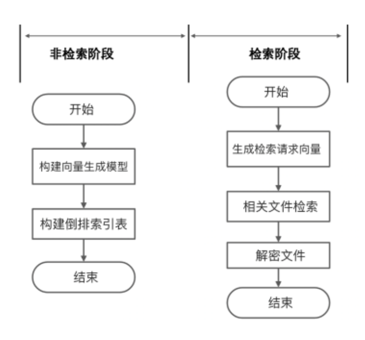


图 4 方案流程

Fig. 4 Scheme process

检索阶段，数据使用者发送检索请求至拥有神经网络向量生成模型的参数服务器，参数服务器收到检索请求后生成检索向量并发送至数据使用者，数据使用者加密检索向量后发送至公有云，公有云通过加密倒排索引表检索出相关文件，并发送给数据使用者；数据使用者通过密钥解密出目标明文数据。

由于数据的更新，会让神经网络向量生成模型不断更新，检索阶段和非检索阶段会重复发生，但分别独立。

4 安全性分析

首先，在神经网络模型中生成陷门和文档索引，并对其进行降维。文档和查询的内容不能直接反映在向量中。此外，还引入了伪关键字、随机分裂和两个 $(n+u+1)^2$ 加密矩阵。正如文献[19]中所证明的，对手无法构造足够的方程来完整地计算矩阵。因此，本文提出的方案很好地抵抗了已知的密文模型威胁模型。

本文将已知背景知识模型下的安全性归结为了解文档索引和检索陷门之间的内在关系。为了进一步防止好奇的公共云服务器根据已知的背景知识泄露和最小化文档信息，本文动态地改变了陷门的表达。本文使用 S 分割密钥 SK 中的向量。因此，即使用户多次检索同一查询，收到的搜索请求陷门也是不同的。同时，所有向量引入随机数 $\varepsilon^{(j)}$，其值服从均值为 μ'，方差为 $\sigma'=c^2/3$ 的均匀分布 $U(\mu'-c,\mu'+c)$，根据中心极限定理，$\varepsilon^{(j)}$ 服从 $N(\mu,\sigma^2)$，σ 值越高，安全性越高但其检索准确性降低，合适的 σ 值，可有效地抵抗了统计分析的攻击。本文提出的方案对于已知的背景知识威胁模型也是安全的。

5 性能评估

本文使用 PYTHON 语言在 Intel Core CPU 为 2.9GHz、Windows10 服务器、RAM 为 4GB 的计算机上实现了该方案。本文提出的系统的性能与 MRSE 方案[2]、PRSE 方案[12]和 FMRSM 方案[20]的性能进行了对比。在实验中，本文在一个真实的数据集上评估整体性能，该数据集以后称为评估数据集，并且利用 DSSM[18]网络作为联邦学习网络模型。从一年的查询文档日志文件中随机选择 20000 个英语查询样本，模型训练时每次一个检索请求和对应的 4 个非相关文档和 1 个相关文档。本文从检索效率和文档检索精度方面进行了性能分析。每次模拟重复 10 次，并分析和给出平均模拟结果，系统的各个模块及具体实现方案如表 1 所示。

表 1 模块介绍与实现方法

Tab. 1 Introduction and implementation method of each module

实现模块	模块实现方法
私有云联邦学习模块	FedML 联邦学习开源框架
联邦学习模型	DSSM[20]
加密倒排索引表模块	自定义算法
文件检索模块	自定义算法
文件加解密密模块	Crypt 库

本文比较了本文的方案与上述方案(MRSE、FMRSM 和 PRSE)的文档检索效率。如图 5 所示，随着集合中文档数量的增加，所有四种方案的检索时间都会增加。MRSE 的搜索时间随着文档集大小的线性增长而近似线性增长。考虑到公共云服务器需要在搜索阶段扫描所有文档索引，这是合理的。FMRSM 和 PRSE 方案的性能更好，但本文的方案的性能优于上述所有方案。本文方案搜索过程基于加密倒排索引表，且向量维数较低。因此，当文档集具有更多文件时，本文的方案更有效。

无论关键字的数量如何，三个方案：MRSE、PRSE 和 FMRSM 的搜索时间大致保持不变，如图 6 所示。但三种方案的检索时间都远高于本文方案。

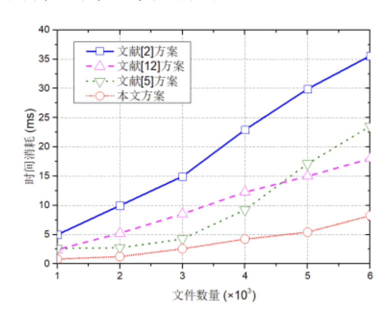


图 5 文件个数对检索时间的影响

Fig. 5 The influence of the number of files on retrieval time

在本文中，本文通过相关文档在所有返回结果中所占的比例来衡量文档检索的准确性。从图 7 可以看出，与 MRSE 相比，无论文档数量如何，本文方案的搜索精度始终高于 95%。相比之下，MRSE 的搜索效率从近 90%逐渐下降到 80%。

图 6 检索关键字对检索时间的影响

Fig. 6 The influence of search keywords on search time

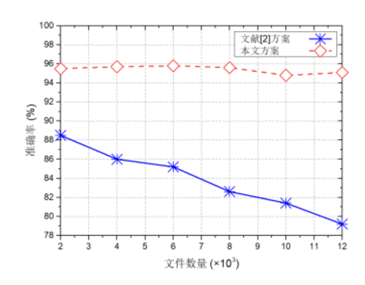


图 7　检索准确率

Fig. 7　Retrieval accuracy

6　结束语

本文提出混合云架构下的深度语义密文检索方案，本文利用私有云进行联邦模型学习，解决了数据孤岛及安全性问题,可提取出文件更深层次语义信息,提高了检索的准确性。并利用加密倒排索引表结构，在保证数据使用者检索关键词不被公有云记录的前提下，提升了检索的效率分析和仿真结果表明，该方案为数据用户提供了安全、高效的加密文档检索服务。

在本文未来的工作中，本文计划优化神经网络结构或使用更好的模型来挖掘加密检索中文档的深层次语义。同时，结合具体模型构建提升检索效率的向量存储模型。

参考文献：

[1] Cao Ning, Wang Cong, Li Ming, *et al.* Privacy-Preserving multi-keyword ranked search over encrypted cloud data [J]. IEEE Transactions on Parallel and Distributed Systems, 2014, 25 (2): 222-233.

[2] Xia, Zhihua, Chen Li, Sun Xingming, *et al.* A multi-keyword ranked search over encrypted cloud data supporting semantic extension [J]. International Journal of Multimedia and Ubiquitous Engineering, 2016, 11. 8: 107-120.

[3] Cai Chengjun, Weng Jian, Yuan Xingliang. *et al.* Enabling reliable keyword search in encrypted decentralized storage with fairness [J]. IEEE Transactions on Dependable and Secure, 2021, 18 (1): 131-144.

[4] Li Feng, Ma Jianfeng, Miao Yinbin, *et al.* Verifiable and dynamic multi-keyword search over encrypted cloud Data Using Bitmap [J]. IEEE Transactions on Cloud Computing. 2021: 1-1.

[5] Liu Lianggui and Chen Qiuxia. A novel feature matching fanked search mechanism over encrypted cloud data [J]. IEEE Access, 2020, 8: 114057-114065.

[6] Zhang Dong, Fan Qing, Qiao Hongyi, *et al.* A public-key encryption with multi-keyword search scheme for cloud-based smart grids [C]// 2021 IEEE Conference on Dependable and Secure Computing, 2021: 1-6.

[7] 沈学利，崔海韵，陈鑫彤．一种支持撤销的位置分层属性加密研究 [J]．计算机应用研究，2019, 37 (1): 216-220. (Shen Xueli, Cui Haiyun, ChenXintong. Research on encryption of location hierarchical attribute ssuporting revocation [J]. Application Research of Computers, 2019, 37 (1): 216-220.)

[8] Miao Yinbin, R. Deng, K. K. R. , *et al.* Threshold multi-keyword search for cloud-based group data sharing [J]. IEEE Transactions on Cloud Computing, 2020, 99: 1-1.

[9] 路宏琳，王利明．面向用户的支持用户掉线的联邦学习数据隐私保护方法 [J]．信息网络安全, 2021, 21 (3): 64-71. (LU Honglin, WANG Liming. User-oriented Data Privacy Preserving Method for Federated Learning that Supports User Disconnection [J]. Netinfo Security, 2021, 21 (3): 64-71.)

[10] 张佳乐，赵彦超，陈兵，等．边缘计算数据安全与隐私保护研究综述 [J]．通信学报, 2018, 39 (3): 1-21. (ZHANG J L, ZHAO Y C, CHEN B, *et al.* Survey on data security and privacy-preserving for the research of edge computing [J]. Journal on Communications, 2018, 39 (3): 1-21.)

[11] Zhang ke, Long Jiahuan, Wang Xiaofen, *et al.* Lightweight searchable encryption protocol for industrial internet of things [J]. IEEE Transactions on Industrial Informatics, 2020, 17 (6): 4248-4259.

[12] Fu Zhangjie, Ren Kui, Shu Jiangang, *et al.* Enabling personalized search over encrypted outsourced data with efficiency improvement [J]. IEEE Transactions on Parallel Distributed Systems, 2016, 27 (9): 2546–2559.

[13] J. Nagesh, N. Jyoti, B. Sayli. Semantic search supporting similarity ranking over encrypted private cloud data [J]. Int. J. Emerging Eng. Res. Technol, 2014, 2 (7): 215–219.

[14] Xia Zhihua, Zhu Yanling, Sun Xingming, *et al.* Secure semantic expansion based search over encrypted cloud data supporting similarity ranking [J]. Journal of Cloud Computing. 2014, 3 (1): 1–11.

[15] Fu Zhangjie, Wu Xinle, Guan Chaowen, *et al.* Toward efficient multi-keyword fuzzy search over encrypted outsourced data with accuracy improvement [J]. IEEE Transactions on Information Forensics and Security. 2017, 11 (12): 2706– 2716.

[16] Yang Wenyuan, Zhu Yuesheng. A verifiable semantic searching scheme by optimal matching over encrypted data in public cloud [J]. IEEE Transactions on Information Forensics and Security. 2020, 16: 100–115.

[17] Wu Wentai, He Ligang, Lin Weiwei, *et al.* SAFA: a semi-asynchronous protocol for fast federated learning with low overhead [J]. IEEE Transactions on Computers. 2020: 1-1.

[18] Shen Yelong, He Xiaodong, Gao Jianfeng, *et al.* A latent semantic model with convolutional-pooling structure for information retrieval [C]// Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. New York: ACM, 2014: 101-110.

[19] Chen Chi, Zhu Xiaojie, Shen Peisong, *et al.* An efficient privacy-preserving ranked keyword search method [J]. IEEE Transactions on Parallel and Distributed Systems. 2016, 27 (4): 951-963.

[20] Wong Kit Wong, D. W. Cheung, B. Kao, *et al.* Mamoulis. Secure knn computation on encrypted databases [C]// Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD) . 2009: 139-152.